# Using the Grid for the BABAR experiment

C. Bozzi, T. Adye, D. Andreotti, E. Antonioli, R. Barlow, B. Bense, D. Boutigny, C.A. J. Brew, D. Colling, R.D. Cowles, et al.

# Using the Grid for the BaBar Experiment

C. Bozzi, T. Adye, D. Andreotti, E. Antonioli, R. Barlow, B. Bense, D. Boutigny, C. A. J. Brew, D. Colling, R. D. Cowles, P. Elmer, E. Feltresi, A. Forti, G. Grosdidier, A. Hasan, H. Lacker, E. Luppi, J. Martyniak, A. McNab, A. Petzold, D. A. Smith, J. E. Sundermann, and P. Veronesi

***Abstract*—The BaBar experiment has been taking data since 1999. In 2001 the computing group started to evaluate the possibility to evolve toward a distributed computing model in a grid environment. We built a prototype system, based on the European Data Grid (EDG), to submit full-scale analysis and Monte Carlo simulation jobs. Computing elements, storage elements, and worker nodes have been installed at SLAC and at various European sites. A BaBar virtual organization (VO) and a test replica catalog (RC) are maintained in Manchester, U.K., and the experiment is using three EDG testbed resource brokers in the U.K. and in Italy. First analysis tests were performed under the assumption that a standard BaBar software release was available at the grid target sites, using RC to register information about the executable and the produced n-tuples. Hundreds of analysis jobs accessing either Objectivity or Root data files ran on the grid. We tested the Monte Carlo production using a farm of the INFN-grid testbed customized to install an Objectivity database and run BaBar simulation software. First simulation production tests were performed using standard Job Description Language commands and the output files were written on the closest storage element. A package that can be officially distributed to grid sites not specifically customized for BaBar has been prepared. We are studying the possibility to add a user friendly interface to access grid services for BaBar.**

***Index Terms*—Distributed computing, elementary particles, physics.**

## I. BaBar Computing Framework

THE BaBar experiment [1] at the SLAC PEP-II asymmetric B Factory has been taking data since summer 1999. The main goal of BaBar is the study of the violation of the combined charge-conjugation and parity (CP) symmetry in the beauty quark sector, and the precise measurement of angles and sides of the unitarity triangle of the Cabibbo–Kobayashi–Maskawa mixing matrix.

Manuscript received November 14, 2003; revised May 13, 2004.

C. Bozzi, D. Andreotti, E. Antonioli, E. Luppi, and P. Veronesi are with INFN Sezione di Ferrara, I-44100, Ferrara, Italy.

T. Adye and C. A. J. Brew are with Rutherford Appleton Laboratory, Chilton Didcot, Oxon OX11 0QX, U.K.

R. Barlow, A. Forti, and A. McNab are with Department of Physics Schuster Lab, University of Manchester, Manchester M13 9PL U.K.

B. Bense, R. D. Cowles, A. Hasan, and D. A. Smith are with Stanford Linear Accelerator Center (SLAC), Stanford, CA 94309 USA.

D. Boutigny is with Lab de Phys. des Particules Chemin du Bellevue B.P. 110 F-74941 Annecy-le-Vieux, Cedex France.

D. Colling and J. Martyniak are with Imperial College Blackett Lab, University of London, London, SW7 2AZ U.K.

P. Elmer is with the Department of Physics, Princeton University, Princeton, NJ 08544 USA.

E. Feltresi, H. Lacker, A. Petzold, and J. E. Sundermann are with Technische Universitaet Dresden Inst. f. Kern- u. Teilchenphysik, D-01062 Dresden, Germany.

G. Grosdidier is with LAL, BP 34 F-91898 Orsay Cedex, France.

Digital Object Identifier 10.1109/TNS.2004.835905

BaBar has collected an integrated luminosity of 130 $fb^{-1}$, corresponding to billions of multihadron events. In terms of mass storage, this corresponds to over 1 PB of data. Thousands of CPUs are required to perform the most resource-intensive tasks such as data reduction (skimming) and Monte Carlo (MC) simulation, in order to keep the end-stage analysis by individual physicists to a manageable level. However, the broad physics program of BaBar reflects in a large number of final users, and therefore the need of massive computing power also for the finals stages of data analysis.

The BaBar computing system is largely distributed, and organized in three levels.

*Tier-A* sites are main computing centers for the experiment, containing all or a significant fraction of data, and providing access to every member of the collaboration. They are also a primary source for data distribution to other smaller sites. Tier-A resources, agreed to with Memoranda of Understanding, consist typically of several hundreds of computers and tens to hundreds of TB of storage. Currently, the BaBar Tier-A centers are located at SLAC, CCIN2P3 (France), RAL (U.K.), INFN-Padova (Italy), and FZK/GridKa (Karlsruhe, Germany). The first three Tier-A sites are intended for data analysis, whereas INFN-Padova is specialized for data (re)processing and FZK is intended for grid developments and data reduction procedures (skimming).

*Tier-B* are centers hosting a subset of data, providing access to a smaller fraction of the BaBar collaborators, typically on a national basis. Tier-B sites have not really been deployed, apart from a site hosted at INFN-Rome for Italian collaborators, which will be eventually upgraded and transformed in a Tier-A center accessible to all collaborators.

*Tier-C* corresponds to single BaBar institutions and host reduced datasets for local analysis. Tier-C sites play an important role in the production of simulated events. Roughly speaking, Babar needs about 1.5 billion generic Monte Carlo events per 100 $fb^{-1}$ integrated luminosity. This figure meets the requirement of simulating at least three times the hadronic cross-section, which is about 5 nb, and is larger than any computing site can practically handle. For this reason, a system was devised and deployed which allows the production of Monte Carlo events to be distributed over 25 sites, including all Tier-A centers [2]. User requests are collected and centrally managed by using a Web interface to a relational database. These requests are then organized in allocations of $\sim$2000 events runs, each run requiring about 5 h of computing. Allocations, whose size depend on the size of the production farm (typically, a few million events) are assigned to production sites. Runs are then managed locally at each site, and events are finally imported to SLAC by

Work supported in part by the Department of Energy contract DE-AC02-76SF00515
Presented at the 2003 IEEE Nuclear Science Symposium & Medical Imaging Conference, 10/19/2003 - 10/25/2003, Portland, OR

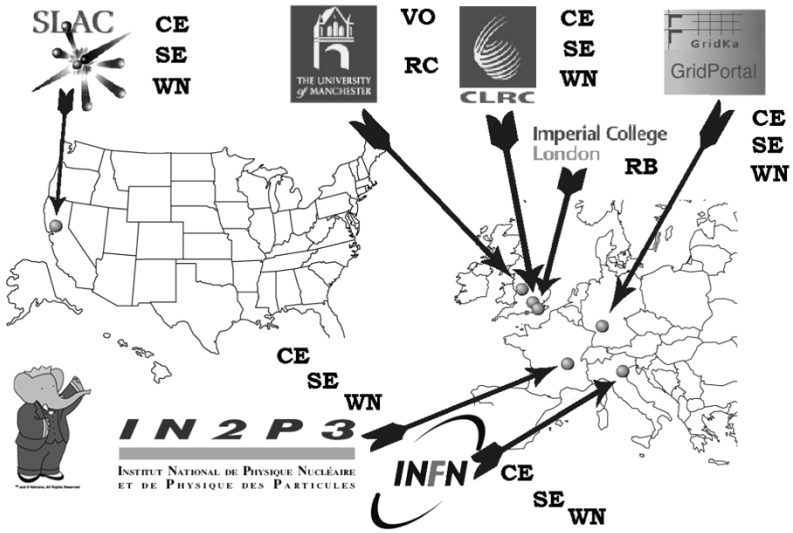


Fig. 1. Configuration of the BaBar grid.

using an automated procedure. The amount of simulated events produced so far is about 3 billion, 80% of which were produced at Tier-C sites.

Two event formats are supported in BaBar, based on the Objectivity object-oriented database system[1] [3] and on ROOT I/O [3], [7], respectively. Because of concerns about scalability and system maintenance, the former is going to disappear as an event store. Nevertheless, the Objectivity database system will be retained to hold detector calibration constants and data taking conditions, due to its good performance in that area.

## II. Motivations for a BaBar Grid

Grid concepts were not taken into account when first deploying the BaBar computing model. However, given the current distributed environment, the grid concept can be very effective in optimizing the usage of resources, and in providing a framework for data analysis where users need not necessarily to know where data are located. For this reason, several grid tools were evaluated.

For data analysis, one has to take into account that data may be spread between several sites, so a metadata catalog is needed to associate the logical name associated to data files with physical locations. Moreover, a simple and reliable tool must be developed in order to automatically split and submit where data and more CPU power are available.

As far as Monte Carlo production is concerned, a grid setup would allow to reduce the manpower needed to run simulation at each of the (currently) 25 production sites. Furthermore, since no bulk input data are needed in Monte Carlo production, one may also consider the possibility to run simulation jobs on grid networks by using non-BaBar resources, when available, thus increasing the production rate and making full use of the Grid paradigm.

Grid tools are also very useful for data distribution among the various computing tiers. Indeed, a Grid tool [4] is now in use to transfer data between SLAC and CCIN2P3.

[1]Objectivity Corp., http://www.objectivity.com.

A driving principle in any evaluation of grid tools is that BaBar is a running experiment producing a major breakthrough in B physics. For this reason, the introduction of any grid technology must not disrupt the existing software and analysis efforts in any way.

## III. The BaBar Grid Infrastructure

The BaBar collaboration has been testing and evaluating mainly the European Data Grid (EDG)[2] [6] middleware as a basis for BaBar grid implementation. This was quite natural since many BaBar participating institutes are also involved in LHC experiments, and some of them are involved in the deployment of the EDG testbed. BaBar is a member of Work Package 8 of EDG, which is the group involved in high energy physics applications of the grid. BaBar is also involved within the U.S. Particle Physics Data Grid,[3] and very active in the development of data storage [4], security, and interactive applications.

Since EDG is now close to an end, BaBar is going to install and test the LHC Computing Grid 1 (LCG1) middleware [5], recently deployed. In particular, national distributions of LCG1, such as Grid.it,[4] are also under evaluation.

The BaBar grid configuration is shown in Fig. 1. The bases are a virtual organization (VO) and a resource broker (RB), maintained in Manchester and Imperial College, U.K., respectively. The VO accepts certificates issued by any EDG certification authority as well as by the U.S. Department Of Energy (DOE). Users with valid certificates copy their headers into a special identification file in the SLAC Andrew file system (AFS) system. A cron job checks for the presence of these files on a regular basis and updates the VO accordingly. The presence of a given set of access control lists (ACLs) on the identification file is the proof that the user is registered in BaBar.

The RB at Imperial College has been regularly upgraded to follow successive EDG releases. Even in its latest versions, the RB software was known to be unstable. The problem was

[2]http://www.edg.org.

[3]http://www.ppdg.net.

[4]http://grid-it.cnaf.infn.it.

that the brokering mechanism relies on a dynamic information system or meta directory service (MDS) that is unable to handle disappearing or reappearing unstable sites. Possible solutions to this problems are discussed in Section IV. Other resource brokers at CNAF (Bologna) and Catania were also used for Monte Carlo production tests.

The other BaBar sites running on EDG testbeds maintain each one or more computing elements (CEs), connected to a farm of worker nodes (WNs) and in some cases to one or more storage elements (SEs). User interfaces (UIs) are used to access grid resources and submit jobs.

A replica catalog, maintained in Manchester, keeps track of the files registered in the SEs. In principle the RC is coupled to a grid data mirroring package (GDMP) system in charge of replicating the files between different SEs. BaBar has deliberately chosen to not install any GDMP system and to manage the replication by hand. It is hoped to use the new replica location service (RLS), announced for version 2.0 of the EDG software. The replica catalog and replica manager (RM) software have several known limitations in term of scalability but their functionalities were nevertheless considered crucial for the studies presented in this work.

SLAC is not part of any EDG testbed infrastructure, so the setup of grid resources needed an ad-hoc customization, to cope in particular with SLAC security requirements, with the SLAC batch scheduler and with the requirement of using only user accounts under AFS. At the moment, SLAC has the EDG1.4 software installed, and will install LCG1.

With respect to the other BaBar grid sites, CCIN2P3 has EDG1.4 and is upgrading to EDG2 and LCG1, Karlsruhe installed LCG1 on UI, CE, SE. Sites in the U.K. installed EDG2, and Imperial College is going to install a LCG1 RB. INFN Ferrara had EDG1.4 and EDG2; LCG1, as distributed in Grid.it, is now installed and running. This allows use also of other Grid.it resources, not necessarily belonging to BaBar. A Web portal (Genius)[5] has been installed in Ferrara for EDG1.4, where a migration to LCG1 is in progress.

## IV. Analysis Jobs

In a typical analysis job, a collection of events is read and either new collections or a set of n-tuples, or Root I/O files, are created.

The user builds the executable locally and uses tool command language (TCL) files provided by the BaBar framework as inputs to select the paths, sequences, and modules which need to be executed at runtime. This usually results in a complicated structure of interdependent TCL files spread around a BaBar software release. It is however possible to produce a complete configuration dump as a single TCL file.

An executable also has other external dependencies on shared libraries (like Objectivity and Root), and on other flat files. All these can be packaged and distributed on the grid, but as a first step, it was assumed that a standard Babar software release is available at grid sites.

A test was performed as follows. First, an executable and an input TCL file were prepared. Since not all the grid farms have a local SE, the executable was then transfered to the SE closest to a set of CE and registered to the RC under a logical file name. A Job Description Language script was prepared in order to send the TCL file and a wrapper script to the CE (through the RB), select a CE close to the SE where the executable is available, run the job on the selected CE, and finally return logfiles to the submission site through the RB.

The wrapper script sets up the directory structure on WN to be compatible with the BaBar setup, copies the executable from the SE to the WN and runs it, transfers the output n-tuple to the SE, and registers it into the RC.

The above test was performed with EDG version 1.4. Several bunches of 200 jobs were submitted, with an executable typical of a real analysis application. As previously said, there are known limitations of the EDG RB software, due to the dynamic MDS not being able to handle disappearing or reappearing unstable sites. The success rate under the default EDG MDS configuration varied from 55% to 75%. To solve this problem, EDG recommends to use the static Berkeley Database Information Index (BDII), which gets updated from MDS every 10 min. If information about a site is not available, the BDII will continue to publish the most recent information it has. By using the static BDII, the success rate went up to 99%. Other inefficiencies due to scalability issues, like the maximum number of jobs present at the same time on the RB, have not been taken into consideration: they will eventually be resolved in future EDG/LCG1 releases.

The focus of current work is to reproduce the above procedure for analysis job submission by using the LCG1 middleware.

## V. Monte Carlo Production

The BaBar Monte Carlo simulation is based on Geant4 [6]. As already explained, events are generated and fully reconstructed in allocations of 2000 event runs by using a system conceptually similar to a computing grid.

With respect to an analysis job, the executable of a simulation job is more stable over time. This suggests a different model for the distribution of Monte Carlo software releases to grid sites. We make use of RedHat package managers RPMs that can be installed at any site running EDG/LCG middleware. These RPMs are packed by using standard system software to determine all calls made by the executable to external shared libraries and TCL or flat files. The objects in the RPM for a typical BaBar software release are the executable, 42 shared libraries (including Geant4 and ROOT), about 600 TCL and flat files, for a total of about 130 MB (36 MB when compressed). Contrary to the case of analysis jobs mentioned above, it is not necessary to have a complete BaBar software release installed at the grid sites. Therefore, the user can run the BaBar Monte Carlo also on non-BaBar resources. Input to the simulation is done through TCL files, and output is produced in form of Root I/O files which are then stored on the closest SE. Data can be then sent back to SLAC or a Tier-A site by using standard EDG commands (edg-copy). Simulation logfiles are returned to the user through the Sandbox mechanism. The input files can be built directly by using the BaBar package for Monte Carlo production

[5]https://genius.ct.infn.it.

(ProdTools) [2], with minimal changes to allow submission to the grid.

A delicate point about Monte Carlo production is that an object-oriented (Objectivity) database is needed to read the detector calibration constants and conditions to be simulated. For this reason, a Sun Ultra-5 machine running the Solaris operating system and sitting outside the grid environment, was configured in Ferrara as a data server which hosts an Objectivity database accessible through the network.

The above scheme for Monte Carlo production was successfully tested by submitting a small number of jobs on EDG1.4.11 in five sites in Italy. All sites needed some manual configuration, for instance to enable access to members of the BaBar VO, and to install the BaBar RPM for Monte Carlo production.

The RPM for Monte Carlo production and the BaBar VO configuration files were then included in the Grid.it software release 1.1.0. Grid.it,[6] is a national Grid project in Italy, with its own testbed and production farms, on which the LCG1 middleware is distributed. Therefore, every site running the Grid.it middleware is able to run the BaBar simulation software without any additional customization.

Tests were then performed by using the CNAF RB, the disk server with the Objectivity database in Ferrara described above, and farms in Ferrara, CNAF, Padova (2), Trieste, INFN-LNL, Bari, Naples, and Catania. Fig. 2 shows the grid infrastructure used for these tests.

Several hundreds of jobs were submitted from the Ferrara UI to the above sites in different times of the day for about one week. Jobs executed in Ferrara had a 95% success rate, whereas jobs submitted to other Italian resources had a success rate of about 60%. Main failures were due to the RB, namely, to Globus services (7%), and to problems in remotely accessing the Objectivity database in Ferrara to read conditions (33%). The latter were due to too many simultaneous connections to the Objectivity database and to network overload (the total bandwidth at Ferrara is 12 Mbit/s), and are not major concerns at the moment. A possible solution is to install a database on a high-performance data server, in a farm where more network bandwidth is available. A complementary approach, also under study, is to install the database on several grid SEs, and to assign a small number of CEs to each SE.

## VI. Evaluation of the Genius Web Portal

Web portals for the grid are designed to allow users to benefit from grid resources without knowing implementation details. The Genius package [7] was tested in BaBar for Monte Carlo production and analysis, and is being evaluated for data analysis. The Genius portal is a Web portal jointly developed by INFN and Nice srl within INFN grid project. It is based on the Engineframe technology. Genius will be also the default portal for generic applications in the project "Enabling Grids for E-science in Europe" (EGEE), funded by the European Union.

A Genius Web server was installed on the INFN Ferrara UI[8] and Genius services for Monte Carlo production were implemented in EDG 1.4.11. It was therefore possible to build and submit Monte Carlo runs, and to retrieve and examine the logfiles, from virtually any browser on any computer.

[6]http://grid-it.cnaf.infn.it.

[7]https://genius.ct.infn.it.

[8]https://grid1.fe.infn.it.

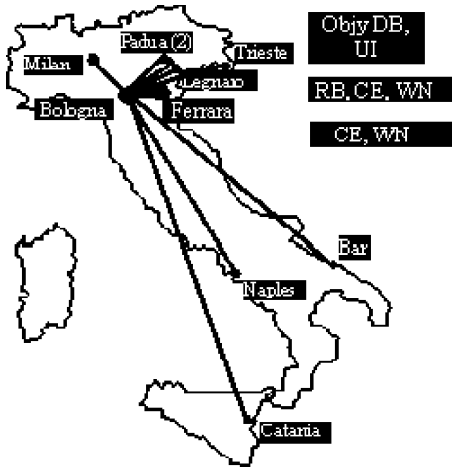


Fig. 2. Grid infrastructure for the Monte Carlo production tests described in the text. The Objectivity database containing detector conditions and backgrounds is located in Ferrara, the RB is at CNAF (Bologna), CE and WN are in Milan, CNAF, Ferrara, Padova (2 sites), Legnaro (INFN-LNL), Trieste, Naples, Bari, Catania. The UI is in Ferrara.

Specifically, the user can get an AFS token, which is mandatory for building simulation runs. Jobs can be created by specifying a set of parameters used by the standard BaBar Monte Carlo simulation, such as run range, number of events per run, detector conditions and backgrounds, which generator to use, any additional filtering instructions, and any flat files containing the decay tree to be generated. An example of how the Web interface can be used for creating and submitting jobs is shown in Fig. 3.

After jobs are submitted, they can be monitored as a function of run range, job identifier, update time, destination CE, and job status, and they may be cancelled during run time. If a run is in the OutputReady status, the output can be retrieved on the UI.

The user can then log in to the UI from the web portal, and examine the logfile and any other files associated to any specific run. The AFS token can be also destroyed at the end, and the user can log off the UI.

As already mentioned, the results of the simulation in the form of Root I/O files are stored on the SE. At the moment, users can transfer these files back to the UI by using EDG commands from the command line. This procedure is currently being implemented in the Web portal, and Genius will be upgraded to use also the LCG1 middleware.

## VII. Outlook

The Monte Carlo production tests outlined in the previous section will be extended to more non-Italian sites (Karlsruhe, Imperial College, Bristol). This will be possible when a resource broker dedicated to BaBar is available. Depending on middleware availability and robustness, it should be possible to use the grid for a substantial amount of Monte Carlo production early next year. Possible problems with reading detector conditions and backgrounds from a single Objectivity database can be overcome by using more network bandwidth and a higher

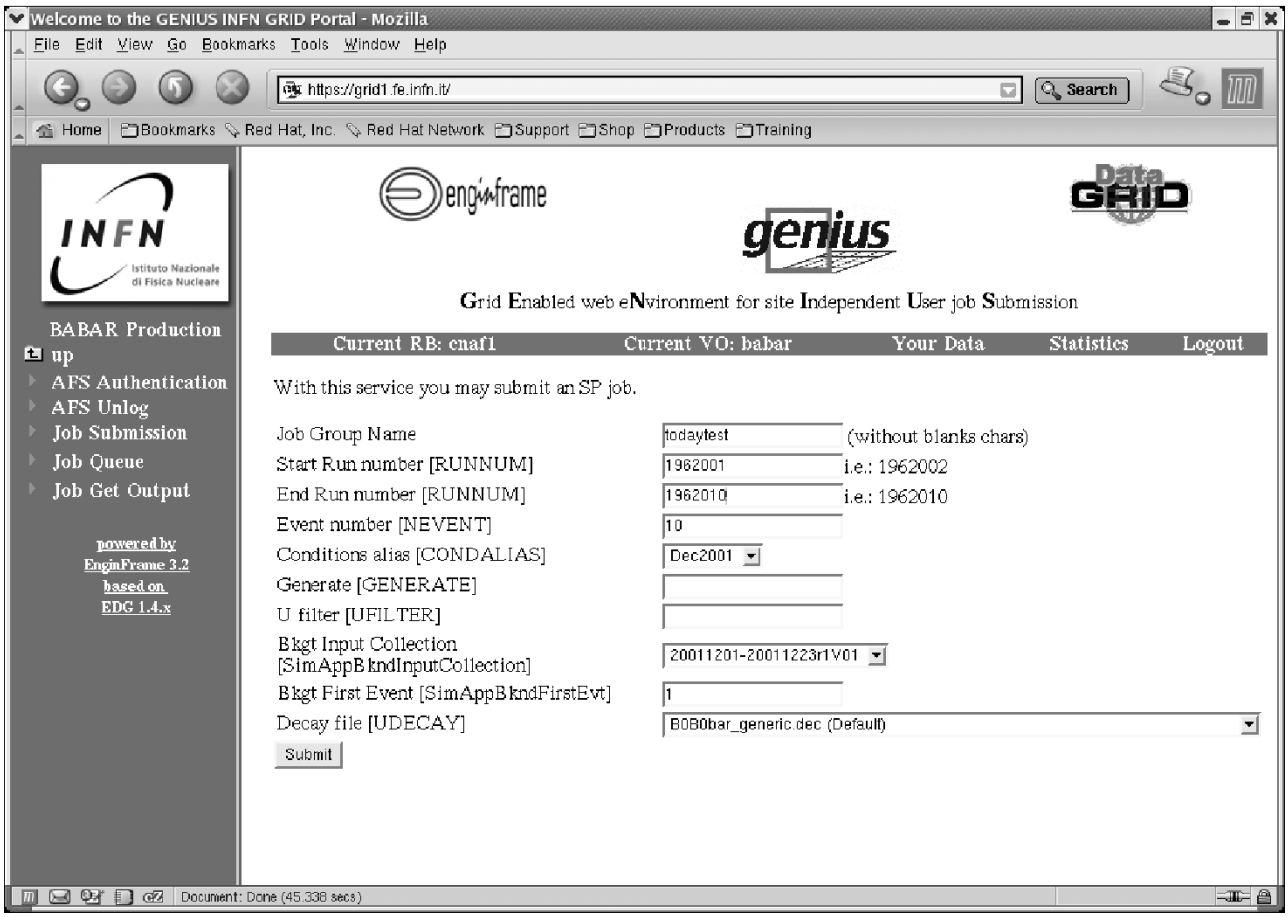


Fig. 3. Example of usage of the Genius portal for BaBar Monte Carlo production. The above screenshot shows a list of parameters (run range, events per run, detector conditions and background, event generator, and filters) which the user can configure to generate simulated events. The resulting jobs are created and submitted to the grid by hitting the "Submit" button at the bottom.

performance data server, and by installing copies of the database in several storage elements.

Grid batch jobs for data analysis should also be possible at some level in early 2004.

A new security method, based on virtual smart cards, will also deployed at SLAC in the next few months.

BaBar grid efforts are migrating from EDG1 and EDG2 to the LCG1 middleware on the same time scale both at SLAC and in other sites.

## VIII. Conclusion

Grid tools have been adapted and tested in the BaBar computing environment, with very encouraging results, compatible with a production system. Grid efforts in Babar are focusing on the support of already existing applications by using underlying existing grid technologies.

For analysis jobs, it is possible to write generic submission scripts and to implement nontrivial tasks on the grid. Some scalability issues remain, but further grid developments are expected to overcome these limitations.

The BaBar simulation software is already available as a package embedded on the Grid middleware, and is therefore distributed and automatically installed on Grid resources. Preliminary tests show that it is possible to manage Monte Carlo production in many remote sites using grid tools and resources. The Genius web portal is a good candidate to embed all user services in a single interface.

The underlying grid framework is still evolving, but it is expected to be reasonably stable in a short time. This will allow the deployment of Monte Carlo production and analysis on grid resources on a timescale of a few months.

## References


[1] B. Aubert, A. Bazan, A. Boucham, D. Boutigny, I. De Bonis, J. Favier, and BaBar Collaboration *et al.*, *Nucl. Instrum. Meth.*, vol. A479, pp. 1–116, 2002.

[2] C. Bozzi, P. Elmer, and D. Smith, "Global management of BaBar simulation production," presented at the Computing in High Energy Physics 2003 Conf., San Diego, CA.

[3] R. Brun and F. Rademaker, "ROOT—an object oriented data analysis framework," in *Proc. AIHENP'96 Workshop*, Lausanne, Switzerland, Sept. 1996.

[4] A. Hasan, W. Kroeger, L. Martin, D. Boutigny, and A. Hanushevsky, "Distributing BaBar data using the storage resource broker (SRB)," presented at the IEEE NSS, Portland, OR, Oct. 2003.

[5] J. Knobloch, "The LHC computing grid project (LCG)," presented at the IEEE NSS, Portland, OR, Oct. 2003.

[6] S. Agostinelli, J. Allison, K. Amako, J. Apostolakis, H. Araujo, and P. Arce *et al.*, "GEANT4: A simulation toolkit," *Nucl. Instrum. Meth.*, vol. A506, pp. 250–303, 2003.

[7] R. Brun and F. Rademaker, "ROOT—an object oriented data analysis framework," *Nucl. Instrum. Meth.*, vol. A389, pp. 81–86, 1997.